Estatística Básica I

REPRESENTAÇÕES GRÁFICAS

TUANY CASTRO

Representações Gráficas

A representação gráfica da distribuição de uma variável tem a vantagem de, rápida e concisamente, informar sobre sua variabilidade.

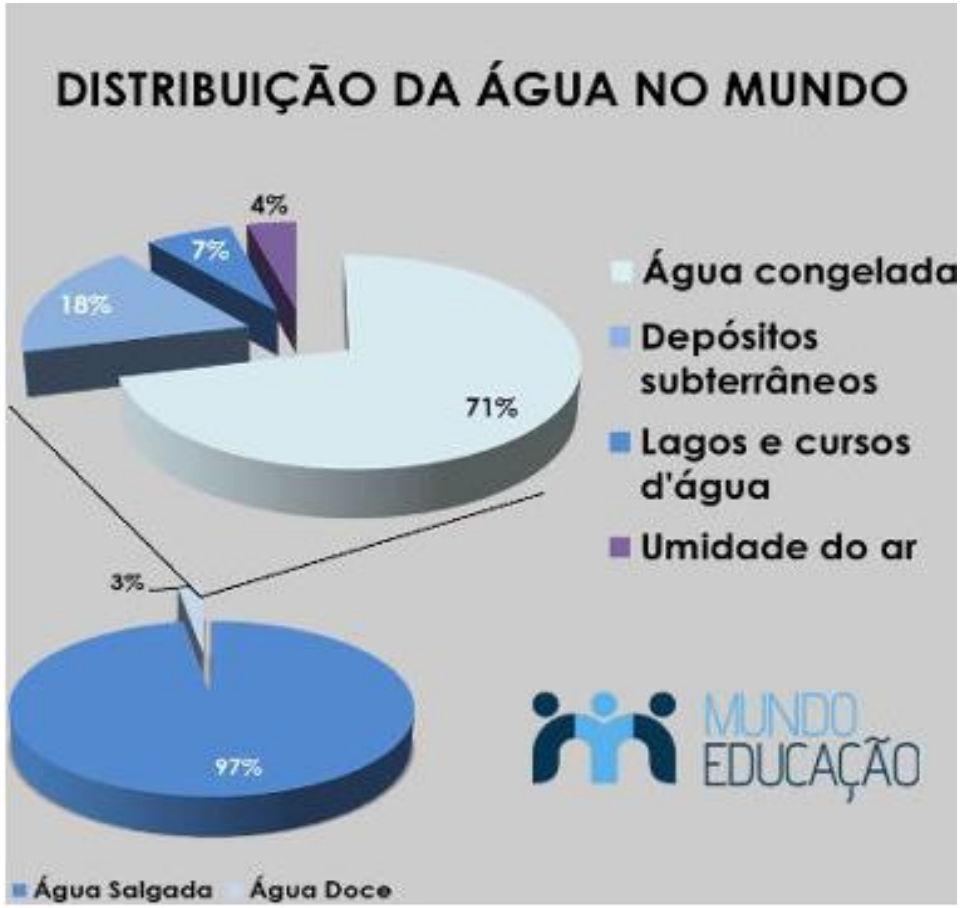
Gráficos para variáveis qualitativas

- Gráficos de setores
- Gráficos de barras

Gráficos para variáveis qualitativas

Gráfico de Setores

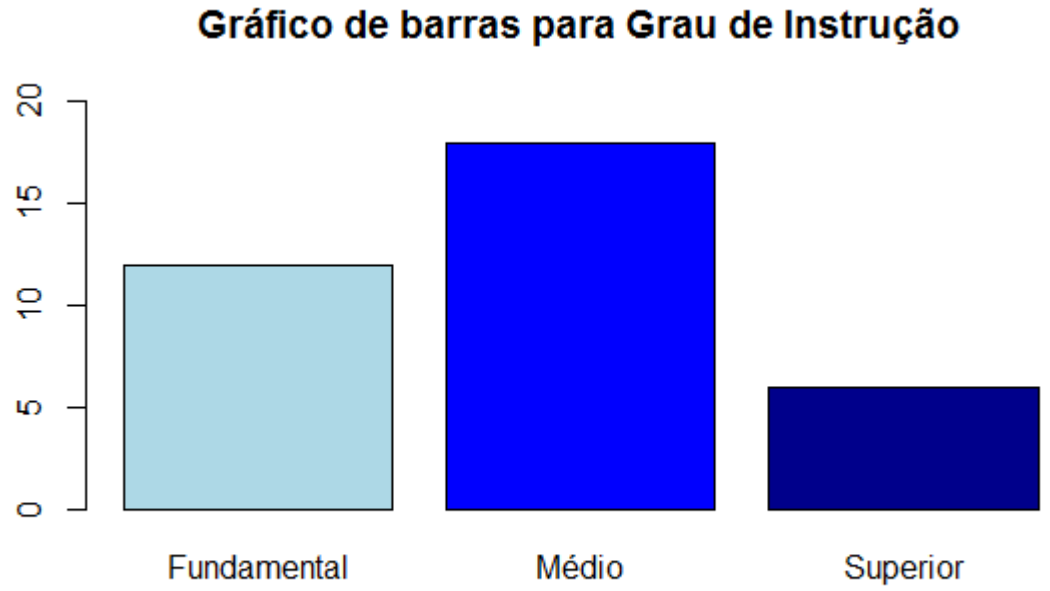
Indicado especialmente para variáveis qualitativas nominais.



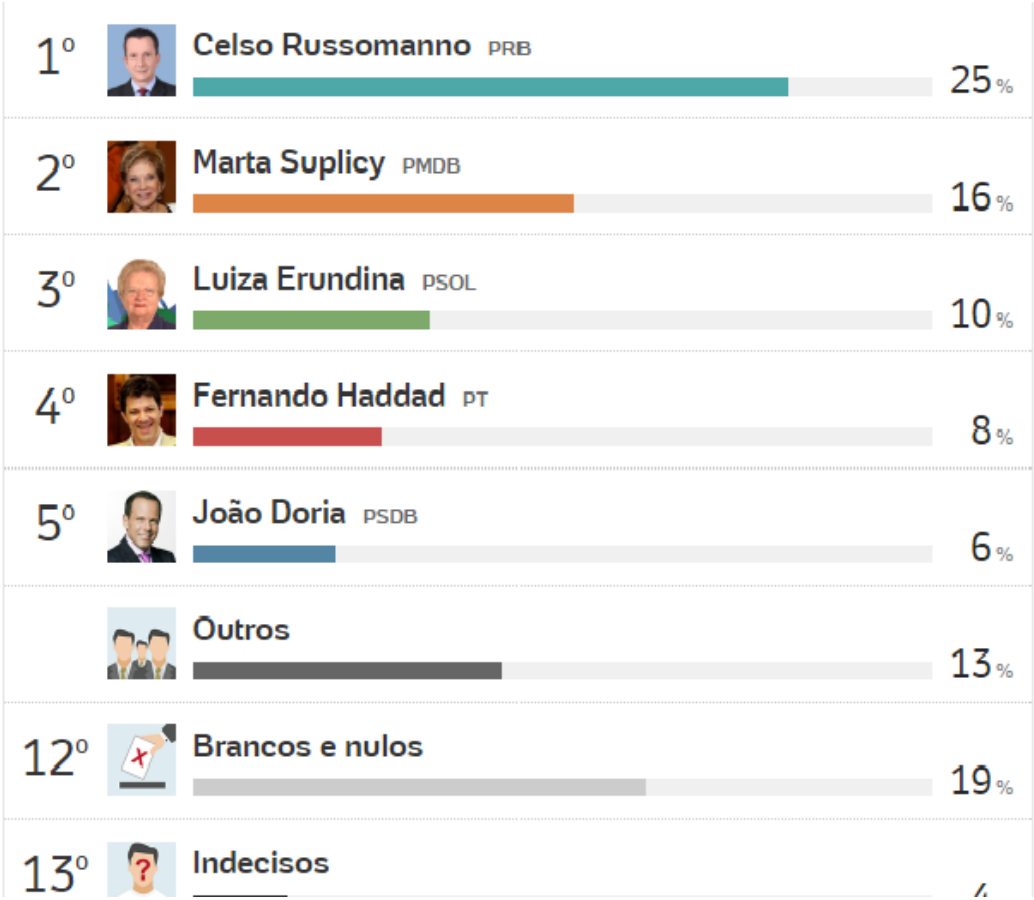
Gráficos para variáveis qualitativas

Gráfico de Barras

Para qualitativas ordinais, o gráfico de barras é mais indicado.



Gráficos para variáveis qualitativas

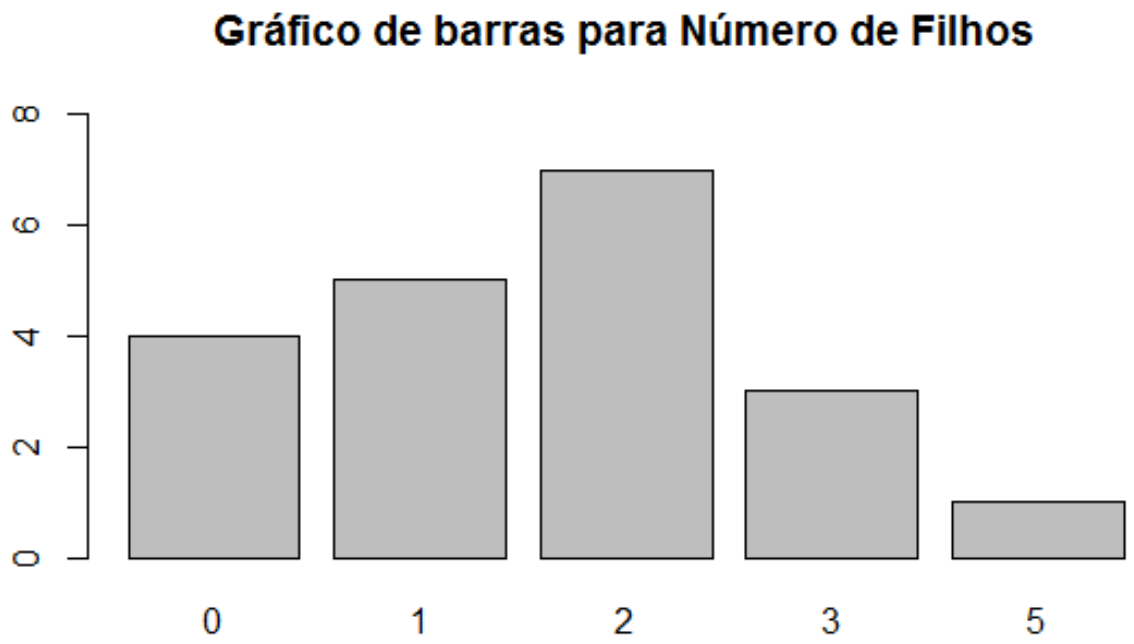


Gráficos para variáveis quantitativas

- Gráfico de barras
- Histograma

Gráfico de Barras

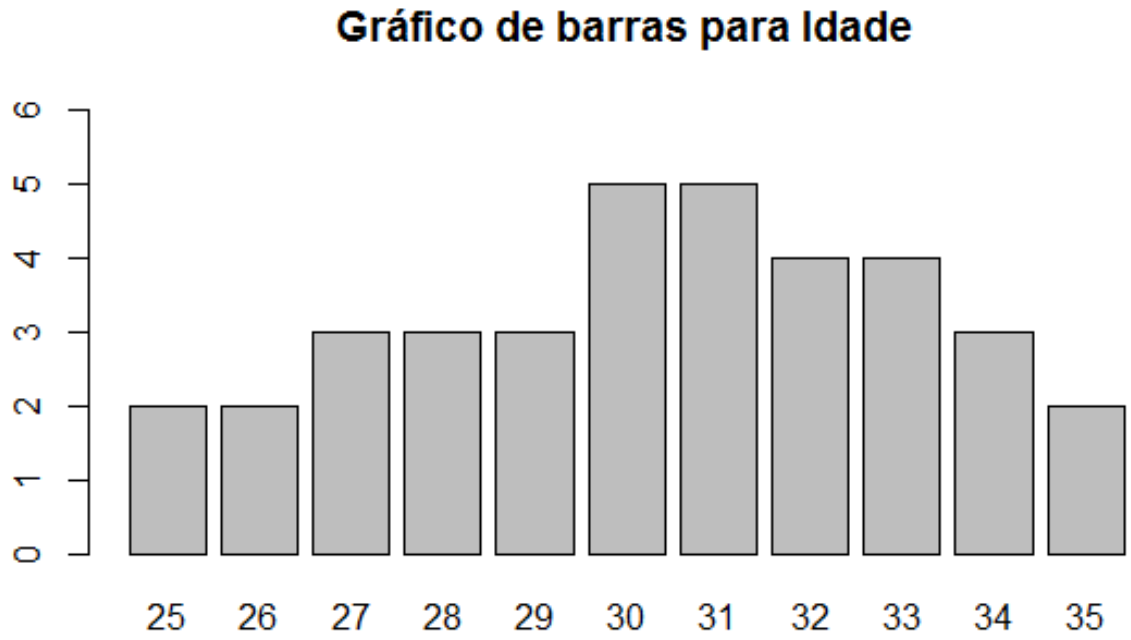
Pode ser utilizado para variáveis quantitativas discretas.



Gráficos para variáveis quantitativas

Gráfico de Barras

Pode ser utilizado para variáveis quantitativas discretas.



Gráficos para variáveis quantitativas

Histograma

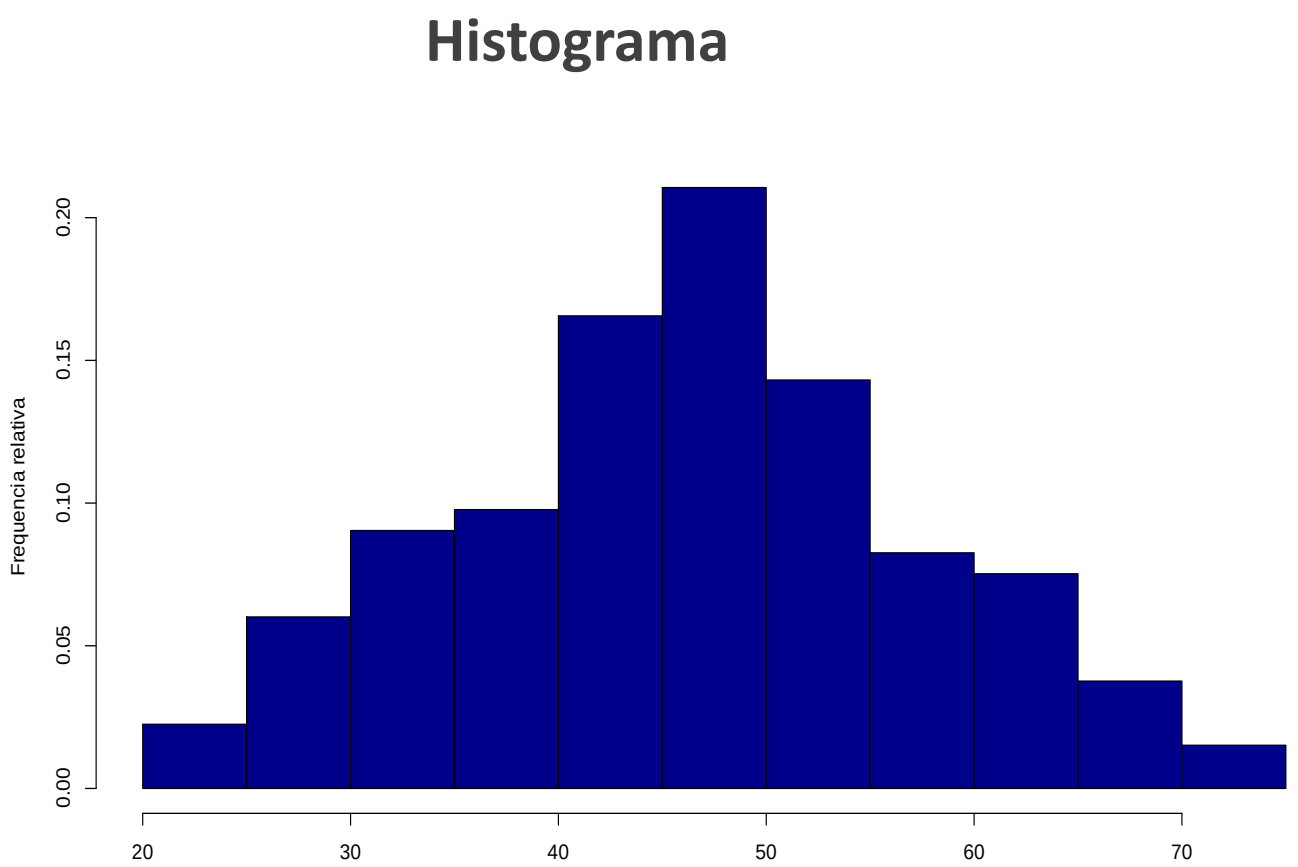
Para variáveis quantitativas contínuas.

- Frequência absoluta do intervalo i: n_i
- Total observado: N
- Frequência relativa do intervalo i: f_i

$$f_i = \frac{n_i}{N}$$

Intervalos	Frequência absoluta	Frequência relativa
20 - 25	3	0,02
25 - 30	8	0,06
30 - 35	12	0,09
35 - 40	13	0,10
40 - 45	22	0,17
45 - 50	28	0,21
50 - 55	19	0,14
55 - 60	11	0,08
60 - 65	10	0,08
65 - 70	5	0,04
70 - 75	2	0,02
Total	133	1,00

Gráficos para variáveis quantitativas



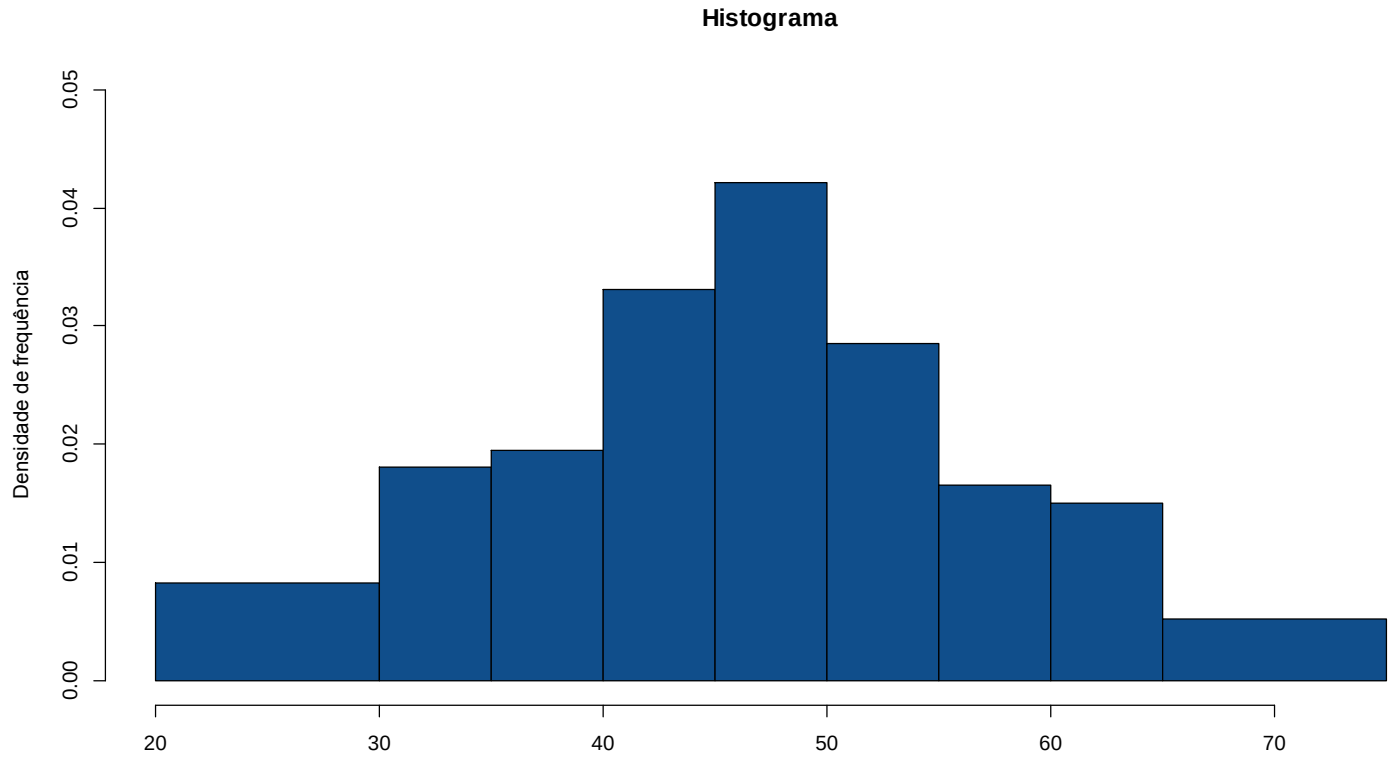
Gráficos para variáveis quantitativas

- Frequência absoluta do intervalo i: n_i
- Total observado: N
- Tamanho do intervalo: t_i
- Frequência relativa do intervalo i: f_i
- Densidade de frequência do intervalo i: d_i

$$d_i = \frac{f_i}{t_i}$$

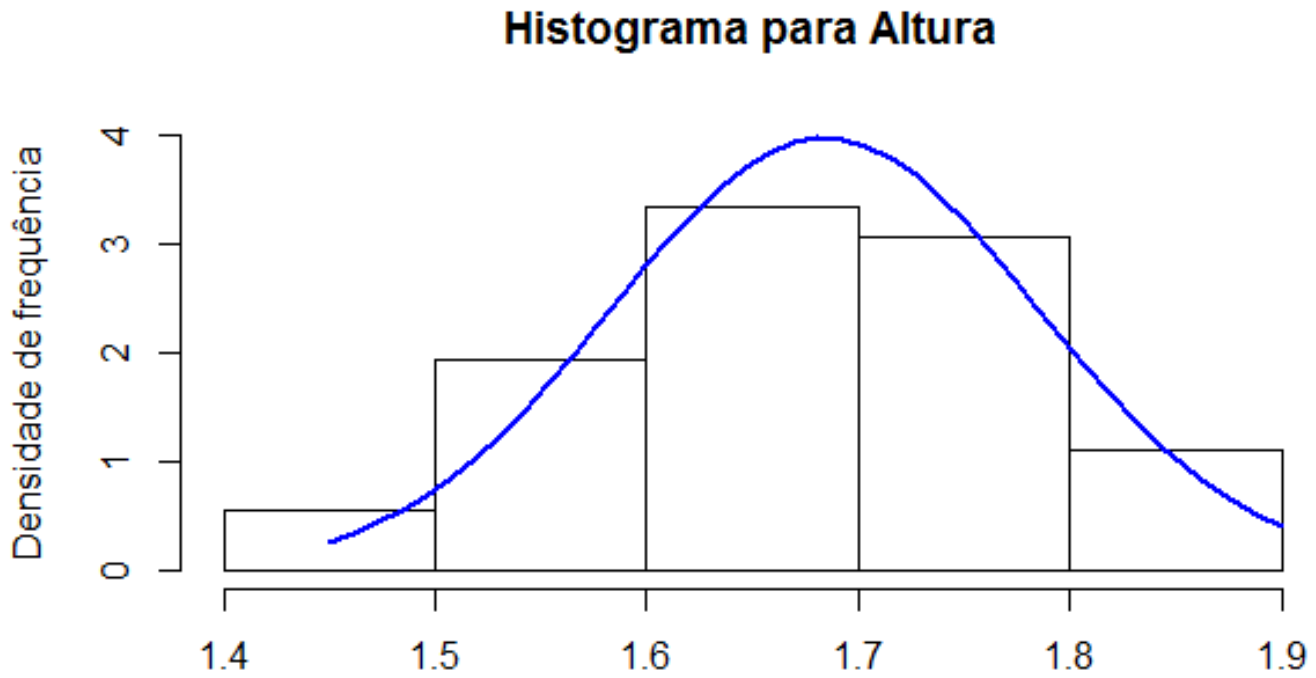
Intervalos	Frequência absoluta	Frequência relativa	Tamanho do intervalo	Densidade de frequência
20 - 30	11	0,08	10	0,01
30 - 35	12	0,09	5	0,02
35 - 40	13	0,10	5	0,02
40 - 45	22	0,17	5	0,03
45 - 50	28	0,21	5	0,04
50 - 55	19	0,14	5	0,03
55 - 60	11	0,08	5	0,02
60 - 65	10	0,08	5	0,02
65 - 75	7	0,05	10	0,01
Total	133	1,00		

Gráficos para variáveis quantitativas



Gráficos para variáveis quantitativas

Histograma alisado

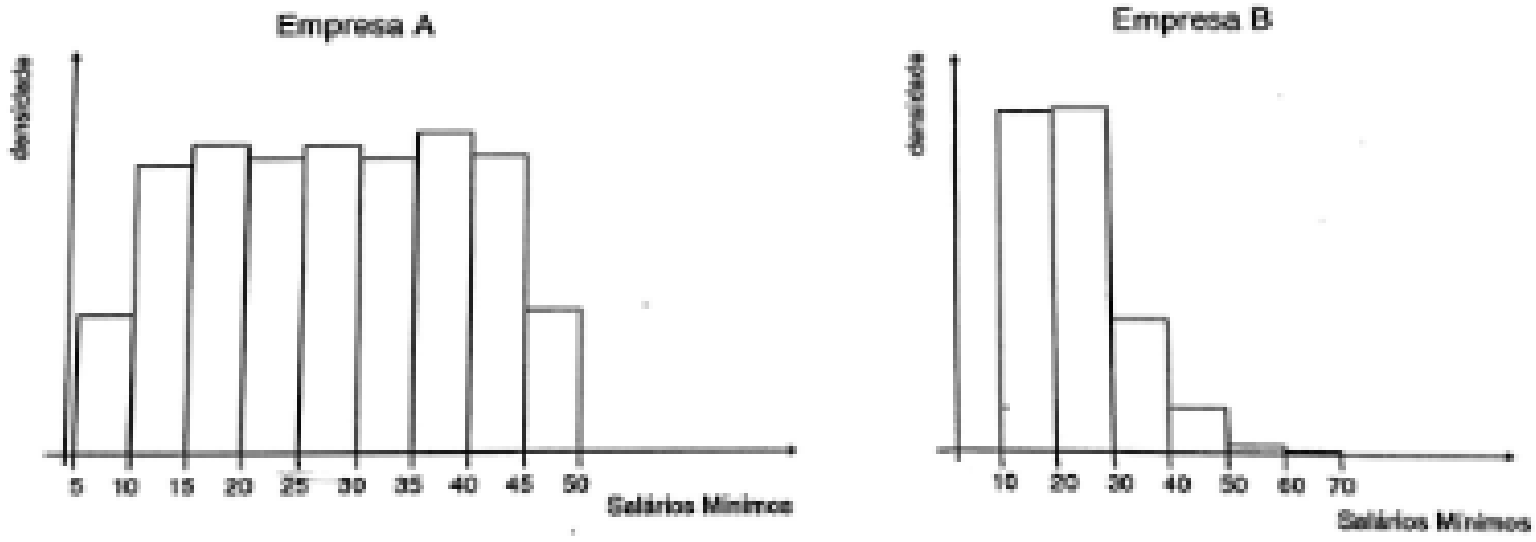


No R: tabelas e medidas resumo

Função	Variável	R
Tabela	Quantitativa discreta ou qualitativa	table(dados$variável) prop.table(table(dados$variável))
Média	Quantitativa	mean(dados$variável, na.rm=TRUE)
Mediana	Quantitativa	median(dados$variável, na.rm=TRUE)
Desvio-Padrão	Quantitativa	sd(dados$variável, na.rm=TRUE)
Medidas Resumo	Quantitativa	summary(dados$variável)

Exercícios

1) Suponha que duas empresas desejam empregá-lo e após considerar as vantagens de cada uma, você vai escolher aquela que lhe pagar melhor. Após certa pesquisa, você consegue a distribuição de salário das empresas, dadas segundo os gráficos abaixo:



Com base nas informações de cada gráfico, qual seria sua decisão?

Exercícios

2) Num experimento, dois grupos de 15 mulheres foram submetidas a dois diferentes tipos de dietas para emagrecimento. Os dados referentes à perda de peso se encontram em 'dados_dieta.csv'. Compare os histogramas dos pesos perdidos das duas dietas.

3) Uma pesquisa com usuários de transporte coletivo na cidade de São Paulo indagou sobre os diferentes tipos usados nas suas locomoções diárias. Dentre ônibus, metrô e trem, o número de diferentes meios de transporte utilizados encontra-se na planilha *tranpostes.csv.*

(a) Construa uma tabela de frequência para a variável Transportes.

(b) Faça uma representação gráfica.

(c) Admitindo que essa amostra represente bem o comportamento do usuário paulistano, você acha que a porcentagem dos usuários que utilizam mais de um tipo de transporte é grande?

Exercícios

4) Os dados do arquivo *comunidade.csv* contém parte dos dados de uma pesquisa sobre aspectos socioeconômicos e culturais de comunidades de baixa renda da região do Butantã em São Paulo. Os dados estão organizados da seguinte forma:

- Coluna 1: número do questionário (Num)
- Coluna 2: Sexo:
  - 1: masculino
  - 2: feminino
- Coluna 3: Idade que começou a trabalhar, em anos

Exercícios

- Colune 4: Série em que parou de estudar
  - N: Não parou de estudar
  - F: parou de estudar no ensino fundamental
  - M: parou de estudar no ensino médio

Caracterize a amostra segundo as variáveis Sexo, Idade e Série por meio de gráficos, tabelas e medidas resumo.

5) No arquivo *temperatura.csv* encontram-se os dados de temperaturas mínimas observadas na cidade de São Paulo em 120 dias de inverno. Construa gráficos adequados para a análise da temperatura.

Exercícios

6) O arquivo *cancer.csv* contém os dados de uma pesquisa sobre incidência de câncer e é apresentado em 4 colunas representando as seguintes variáveis de interesse:

- coluna 1: identificação do paciente
- coluna 2: diagnóstico:
  - 1 = falso-negativo
  - 2 = negativo
  - 3 = positivo
  - 4 = falso-positivo
- coluna 3: idade
- coluna 4: glicose (GL)

Exercícios

(A) Construa gráficos adequados para a análise das variáveis Grupo, Idade e Glicose (GL).

(B) Uma afirmação feita por alguns médicos é a de que o grupo dos falso-positivos é mais jovem do que o dos falso-negativos. Para os dados dessa pesquisa, o que você diria a respeito? Justifique sua resposta baseando-se em gráficos.

Quantis

Medidas de locação calculadas para variáveis quantitativas.

Um quantil de ordem p é o valor tal que p% das observações são menores do que ela.

Alguns quantis têm nomes específicos:

- **Quartis**: quantis de ordem 4, pois dividem os dados em 4 intervalos
- **Percentis**: quantis de ordem 100, pois dividem os dados em 100 intervalos

Quartis

Assim como a mediana que divide o conjunto de dados em duas metades, os quartis dividem este conjunto em quartos.

1° Quartil – é o valor que deixa 25 % dos dados abaixo e 75% acima dele.

2° Quartil – Mediana.

3° Quartil – é o valor que deixa 25% dos dados acima e 75% abaixo dele.

Percentis

Os percentis dividem o conjunto de dados em 100 partes.

1° Percentil – é o valor que deixa 1 % dos dados abaixo e 99% acima dele.

2° Percentil – é o valor que deixa 2 % dos dados abaixo e 98% acima dele.

⋮

50° Percentil – mediana.

⋮

99° Percentil – é o valor que deixa 99 % dos dados abaixo e 1% acima dele.

.

Cálculo de quartis e percentis no R

- **Medidas de posição (média, mediana, primeiro e terceiro quartis):**

```
summary(dados$variável)
```

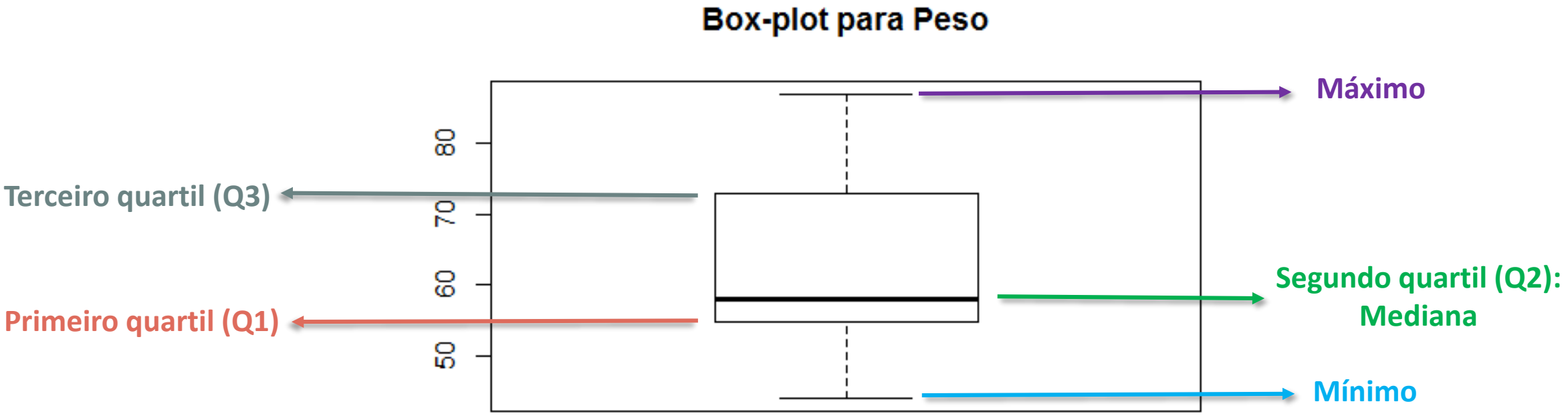
- **Quartis:**

```
quantile(dados$variável, probs = seq(0, 1, 0.25), na.rm=TRUE)
```

- **Percentis:**

```
quantile(dados$variável, probs = seq(0, 1, 0.01), na.rm=TRUE)
```

Box-Plot



Representação gráfica envolvendo quartis.

Box-Plot

- Possibilidade observação da **variabilidade** e da **simetria** dos dados.

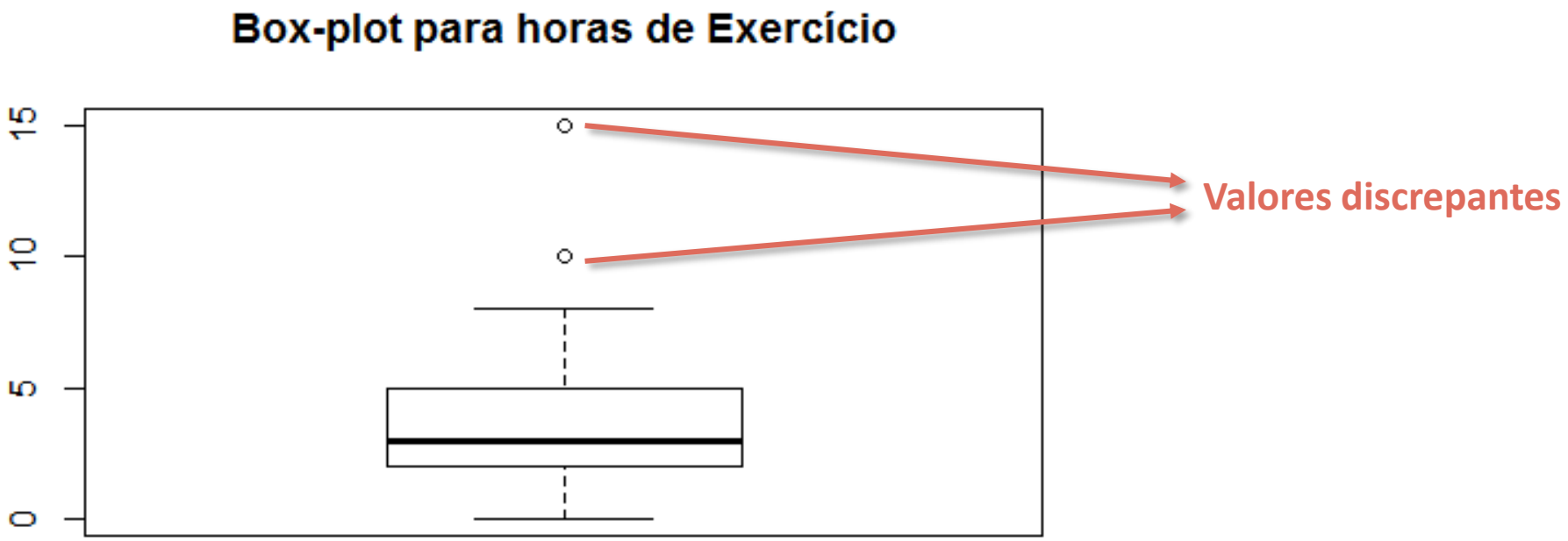
Distribuição simétrica

Distâncias iguais da mediana para os quartis

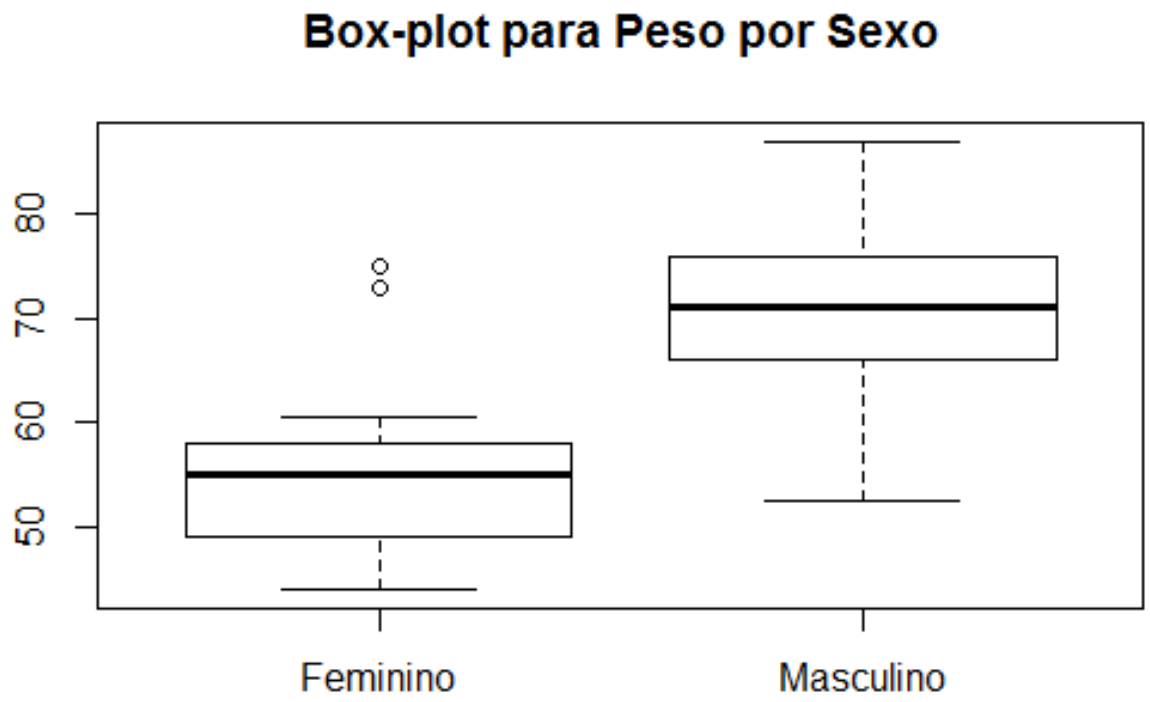
Distâncias iguais dos pontos de mínimo e máximo em relação à mediana

- **Identificação de valores discrepantes**
  - valores acima de Q3 +1,5*(Q3-Q1)
  - valores abaixo de Q1-1,5*(Q3-Q1)

Box-Plot


Box-plot para horas de Exercício
15
10
5
0
Valores discrepantes

Comparação de Box-Plot



Box-Plot no R

- **Box-Plot para uma variável:**

boxplot(dados$variável, main = “Título do gráfico”

- **Box-Plot de uma variável dentro dos grupos de outra variável:**

boxplot(dados$variável ~ dados$grupo, main = “Título do Gráfico”, names = c(nomes das categorias)

Exercícios

O arquivo *cancer.csv* contém os dados de uma pesquisa sobre incidência de câncer e é apresentado em 4 colunas representando as seguintes variáveis de interesse:

- coluna 1: identificação do paciente
- coluna 2: diagnóstico:
  - 1 = falso-negativo
  - 2 = negativo
  - 3 = positivo
  - 4 = falso-positivo
- coluna 3: idade
- coluna 4: glicose (GL)

Compare as distribuições de Glicose entre os grupos por meio do gráfico box-plot.

Exercício no R

Os dados em *estudo_cardio.xlsx* são provenientes de um estudo sobre teste de esforço cardiopulmonar em pacientes com insuficiência cardíaca realizado no InCor. As variáveis do estudo são:

- IDENTIFICACAO: identificação do paciente
- ETIOLOGIA: CH – chagásicos, ID – idiopáticos, IS – isquêmicos, C - controle
- SEXO: F – feminino, M - masculino
- IDADE: em anos
- IMC: índice de massa corpórea (kg/m2)
- FC: frequência cardíaca em batimentos por minuto
- VO2: consumo de oxigênio em ml/(kg.min).

Exercício no R

A) Construa uma tabela de frequências e um gráfico de setores para a variável sexo.

B) Construa uma tabela de frequências e um gráfico de barras para a variável etiologia.

C) Calcule medidas resumo e construa histogramas para as variáveis IMC e FC.

D) Calcule medidas resumo e construa um boxplot para VO2.

E) Calcule medidas resumo e construa um boxplot para VO2 por sexo. Você acha que há diferença?

F) Calcule medidas resumo e construa um boxplot para VO2 por etiologia. Você acha que há diferença?

G) Construa um gráfico de dispersão de VO2 por FC. Há alguma tendência?

H) Construa um gráfico de dispersão de VO2 por IMC. Há alguma tendência?

I) O que pode se concluir com essa análise descritiva?